Introduction to Maximum Likelihood Estimation

S. Gliske, University of Michigan

August 7, 2009

Abstract

Maximum Likelihood Estimation (MLE) is a one of the best and most standard methods of density estimation. The purpose of this document is to collect various generalities regarding MLE relevant to HERMES analysis into one document. Specific items regarding interpretation are included in other documents.

1 Differential Yields and Probability Densities

A differential yield over D-dimensions can be defined as members of the set $\mathcal{F}_D$, the set of all functions that map a subdomain $\mathcal{D}$ of a D-dimensional real space to non-negative finite real numbers. A probability density function (PDF) $p(\boldsymbol{x})$, $x \in \mathbb{R}^D$, is a normalized yield,

$$\int_{\mathcal{D}} d^D\boldsymbol{x}\; p(\boldsymbol{x}) = 1, \tag{1}$$

whose range is finite. If p has a parametric form depending on parameters $\boldsymbol{\alpha}$, then we write $p(\boldsymbol{x}|\boldsymbol{\alpha})$. This is read as the probability density at $\boldsymbol{x}$ given $\boldsymbol{\alpha}$. Such a quantity is a conditional probability density.

2 Maximum Likelihood Estimation

Given a parametric conditional probability $p(\boldsymbol{x}|\boldsymbol{\alpha})$ and a data set $\{\boldsymbol{x}^{(i)}\}_{i=1}^{n}\}$, the likelihood of $\boldsymbol{\alpha}$ given the data is

$$L(\boldsymbol{\alpha}) = \prod_{i=1}^{n} p(\boldsymbol{\alpha}|\boldsymbol{x}^{(i)}). \tag{2}$$

This can be re-expressed in terms of our chosen $p(\boldsymbol{x}|\boldsymbol{\alpha})$ by using Bayes rule, yielding

$$L(\boldsymbol{\alpha}) = \prod_{i=1}^{n} \frac{p(\boldsymbol{\alpha})}{p(\boldsymbol{x}^{(i)})} p(\alpha|\boldsymbol{x}^{(i)}). \tag{3}$$

The PDFs $p(\boldsymbol{\alpha})$ and $p(\boldsymbol{x}^{(i)})$ are called the priors, and they represent the assumptions about $\boldsymbol{\alpha}$ and $\boldsymbol{x}$ prior considering them jointly, i.e. the distributions independent of each other. As we make no assumptions about $\boldsymbol{\alpha}$ that are not based on the data, and as we assume $\boldsymbol{x}$ has no further structure than that parametrized by $\boldsymbol{\alpha}$, both of these terms are constant. Thus we have

$$L(\boldsymbol{\alpha}) \propto \prod_{i=1}^{n} p(\boldsymbol{x}^{(i)}|\boldsymbol{\alpha}). \tag{4}$$

It is numerically preferable to work with the logarithm of this expression, the log-likelihood

$$LL(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \ln p(\boldsymbol{x}^{(i)}|\boldsymbol{\alpha}) + \text{const.} \tag{5}$$

MLE is a method of determining the optimal parameters, given a data set and a parametric form for $p(\boldsymbol{x}^{(i)}|\boldsymbol{\alpha})$, and is based on the assumption that the optimal parameters are those parameters that are most likely given the data, i.e. that maximize the above likelihood (or log-likelihood) function. This is usually written

$$\widehat{\boldsymbol{\alpha}} = \arg_{\boldsymbol{\alpha}} \max \sum_{i=1}^{n} \ln p(\boldsymbol{x}^{(i)}|\boldsymbol{\alpha}), \tag{6}$$

where $\widehat{\boldsymbol{\alpha}}$ is the MLE estimate of the parameters. Numerically, one can equivalently minimize the negative of the above expression, although in speaking and writing one usually always refers to the log-likelihood and maximizing, rarely the negative log-likelihood and minimizing.

3 MLE with Yields

A differential yield is related to a PDF by dividing by the integral,

$$p(\boldsymbol{x}) = \left[\int_{\mathcal{D}} d^D\boldsymbol{x}\ f(\boldsymbol{x}) \right]^{-1} f(\boldsymbol{x}). \tag{7}$$

and similarly with conditional probability densities. The integral $\int_{\mathcal{D}} d^D\boldsymbol{x}\ f(\boldsymbol{x})$ is the expected total yield, i.e. number of events.

This can be substituted into the log-likelihood function. Note the integral does not depend on the data, and so can be taken out of the sum. The MLE estimate with a non-normalized function then has the form

$$\widehat{\boldsymbol{\alpha}} = \arg_{\boldsymbol{\alpha}} \max \left[\sum_{i=1}^{n} \ln f(\boldsymbol{x}^{(i)}, \boldsymbol{\alpha}) - n \ln \int_{\mathcal{D}} d^D\boldsymbol{x}'\ f(\boldsymbol{x}', \boldsymbol{\alpha}) \right], \tag{8}$$

where the prime has been added to the integration variable $\boldsymbol{x}'$ to further distinguish it from the data variables $\boldsymbol{x}^{(i)}$.

The integral can either be computed analytically or numerically. A typical method of numeric integration in several dimensions is called Monte Carlo integration. In this case, one chooses a set $\{\boldsymbol{x}'^{(j)}\}_{j=0}^{m}$ which is distributed uniformly at random within the domain $\mathcal{D}$. The integral is then approximated by

$$\int_{\mathcal{D}} d^D\boldsymbol{x}'\ f(\boldsymbol{x}', \boldsymbol{\alpha}) \approx \frac{V_{\mathcal{D}}}{m} \sum_{j=0}^{m} f(\boldsymbol{x}'^{(j)}, \boldsymbol{\alpha}), \tag{9}$$

where $V_{\mathcal{D}}$ is the volume of the domain. As only the logarithm of this integral occurs in the log-likelihood, the factor $V_{\mathcal{D}}/m$ contributes a term constant with respect to $\boldsymbol{\alpha}$ and can be ignored.

Note that if the set $\{\boldsymbol{x}'^{(j)}\}_{j=0}^{m}$ is not chosen uniformly at random, but according do some distribution $p'(\boldsymbol{x})$, then one actually has approximated

$$\frac{1}{m} \sum_{j=0}^{m} f(\boldsymbol{x}'^{(j)}, \boldsymbol{\alpha}) \approx \int_{\mathcal{D}} d^D\boldsymbol{x}'\ p'(\boldsymbol{x}) f(\boldsymbol{x}', \boldsymbol{\alpha}). \tag{10}$$

This fact is used to relate the Kullback-Leibler divergence to MLE, extending the use of MLE to other fields and extending its interpretation. This fact is will also be helpful when including the acceptance, Section 5.

4 Uncertainty

A somewhat approachable treatment of uncertainty calculations and confidence levels can be found in Reference [1], mainly based on the book in Reference [2]. The important result in our case is that the covariance matrix C for the parameters $\boldsymbol{\alpha}$ is the negative inverse Hessian matrix for the log-likelihood,

$$\left(C^{-1}\right)_{i,j} = -\frac{\partial^2}{\partial\alpha_i\partial\alpha_j} LL\bigg|_{\boldsymbol{\alpha}=\widehat{\boldsymbol{\alpha}}}. \tag{11}$$

The negative can be thought of as arising from the fact that one is maximizing instead of minimizing. Alternately, one can state the covariance matrix is the Hessian matrix of the negative log-likelihood.

5 Acceptance

In any real experiment, both acceptance and smearing affects are present. Thus rather than having simply the true distribution $p_T(\boldsymbol{x}_T)$, depending on the true values of the variables $\boldsymbol{x}^T$, one instead has the measured distribution $p_M(\boldsymbol{x}_M)$, related by some conditional probability $p(\boldsymbol{x}_M|\boldsymbol{x}_T)$, the differential probability of measuring $\boldsymbol{x}_M$ given that $\boldsymbol{x}_T$ really occurred. The expression is a Fredholm integral equation,

$$p_M(\boldsymbol{x}_M) = \int_{\mathcal{D}_T} d\boldsymbol{x}_T\; p(\boldsymbol{x}_M|\boldsymbol{x}_T)\, p_T(\boldsymbol{x}_T), \tag{12}$$

and is further treated in Hermes Internal Note 08-015 [3], among other places [4].

Assuming smearing effects are small is assuming the conditional probability $p(\boldsymbol{x}_M, \boldsymbol{x}_T)$ is approximately proportional to $\delta^D(\boldsymbol{x}_M - \boldsymbol{x}_T)\epsilon(\boldsymbol{x}_M)$, or that the measured distribution is

$$p_M(\boldsymbol{x}) \propto \epsilon(\boldsymbol{x}) p_T(\boldsymbol{x}). \tag{13}$$

Note the proportional sign, since ϵ changes the normalization.

5.1 Correcting with Normalization Monte Carlo[1]

To account for acceptance (neglecting smearing), one can follow the method presented by Andy Miller [6], repeated here with slight typographical changes. Factor $p_M(\boldsymbol{x})$ into the angular integrated unpolarized cross-section $\underline{\sigma}_{UU}$ and remaining portion, g. Let $\boldsymbol{y}$ denote the non-angular variables and $\boldsymbol{\phi}$ the angular variables. One would then like to use the non-normalized function f,

$$f(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\alpha}) = \epsilon(\boldsymbol{y}, \boldsymbol{\phi}) \underline{\sigma}_{UU}(\boldsymbol{y}) g(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\alpha}) \tag{14}$$

to find the MLE estimate of the parameters $\boldsymbol{\alpha}$. Note, in taking the logarithm of f, the factors $\epsilon(\boldsymbol{y}, \boldsymbol{\phi})$ just contribute a constant offset to the log-likelihood and thus can be ignored. However, this is not true for the normalization integral,

$$\int_{\mathcal{D}} d\boldsymbol{y} d\boldsymbol{x}\; f(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\alpha}) = \int_{\mathcal{D}} d\boldsymbol{y} d\boldsymbol{x}\; \epsilon(\boldsymbol{y}, \boldsymbol{\phi}) \underline{\sigma}_{UU}(\boldsymbol{y})\; g(\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\alpha}), \tag{15}$$

as the logarithm is taken after the integral.

[1] This section describes the rational behind this method. The applicability and interpretation of this method is discussed in Reference [5]. Do not use this method until reading this other document.

Although the product $\epsilon(\boldsymbol{y},\boldsymbol{\phi})\underline{\sigma}_{UU}(\boldsymbol{y})$ is unknown, Monte Carlo data can be generated according to this distribution. Thus using Monte Carlo integration and Monte Carlo data, one has

$$\int_{\mathcal{D}} d\boldsymbol{y}d\boldsymbol{x}\; f(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{\alpha}) \;\;\approx\;\; \frac{1}{m}\sum_{j=1}^{m} g(\boldsymbol{y}^{(j)},\boldsymbol{\phi}^{(j)},\boldsymbol{\alpha}), \tag{16}$$

where m is the number of Monte Carlo data points used. Putting this into the MLE expression, one finally has

$$\widehat{\boldsymbol{\alpha}} = \arg_{\boldsymbol{\alpha}} \max \left[\sum_{i=1}^{n} \ln g(\boldsymbol{y}^{(i)},\boldsymbol{\phi}^{(i)},\boldsymbol{\alpha}) - n \ln \sum_{j=1}^{m} g(\boldsymbol{y}^{(j)},\boldsymbol{\phi}^{(j)},\boldsymbol{\alpha}) \right], \tag{17}$$

recalling the sum over i is using real data, while the sum over j is using Monte Carlo data. If, instead, one mistakenly uses the real data in both sums, the result approximates

$$\sum_{i=1}^{m} g(\boldsymbol{y}^{(i)},\boldsymbol{\phi}^{(i)},\boldsymbol{\alpha}) \approx m \left[\int_{\mathcal{D}} d\boldsymbol{y}d\boldsymbol{\phi}\; g^2(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{\alpha}) \right] \left[\int_{\mathcal{D}} d\boldsymbol{y}d\boldsymbol{\phi}\; g(\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{\alpha}) \right]^{-1}, \tag{18}$$

a quantity not useful in this context, rather than approximating the needed normalization integral.

References

[1] Geyer, C.J. Class Notes. `http://www.stat.umn.edu/geyer/5102/notes/fish.pdf`

[2] Degroot, M. and Schervish, M. *Probability and Statistics*, Addison-Wesley Publishing, 2002 (ISBN: 0201524880).

[3] Hermes Internal Note 08-015 discusses the mathematics of unfolding and presents unfolding methods using histograms and KDEs.

[4] Additonal references and current information is available at
`http://hermes-wiki.desy.de/index.php/Unfolding_of_Detector_Smearing_and_QED_Radiative_Effects`

[5] Gliske, S. "Note Regarding MLE and Asymmetries."
`http://www-hermes.desy.de/PLOTS/0908/sgliske/Gliske.MLE_Interpretation.pdf`

[6] Andy Miller has discussed the matter on several occasions, including

`http://www-hermes.desy.de/groups/mgmtgrp/COLLABMEETINGS/TRANSVERSITY_JUNE06/MaxLike.pdf`
`http://www-hermes.desy.de/groups/mgmtgrp/COLLABMEETINGS/OCT05/TRANSVERSITY/Andy_unbinned.pdf`
`http://www-hermes.desy.de/groups/mgmtgrp/COLLABMEETINGS/DEC05/TRANSVERSITY/Andy_MaxLike.pdf`